Chapter 21

Superalignment and Safety Scaling Law in AI Agents

21.1 Superalignment: Goal-Driven Alignment for AI Agents

As LLMs increasingly serve as the core of decision making of autonomous agents, ensuring that their output remains safe, ethical, and consistently aligned with human objectives has become a pressing challenge [1386, 402, 1387]. Traditional alignment techniques, particularly RLHF, have been instrumental in refining LLM behavior by incorporating human preferences [110, 43].

Traditional safety alignment focuses primarily on preventing harmful outcomes by enforcing predefined constraints. In such frameworks, an agent's behavior is guided by a single aggregated reward signal that prioritizes immediate corrections over long-range planning. Although this reactive approach works in many current applications, it struggles when an agent must execute extended, multifaceted tasks. The inability to decompose intricate, long-term goals into interpretable and manageable sub-objectives may result in behavior that is technically safe yet suboptimal for fulfilling broader human-centric aims.

To address these limitations, the concept of **superalignment** [1388] has emerged. Superalignment represents an evolution in alignment strategies by embedding explicit long-term goal representations directly into an agent's decision-making process. Rather than simply imposing constraints to avoid harmful actions, superalignment proactively governs behavior through a composite objective function. This function integrates several dimensions of performance—specifically, safety and ethical considerations (where ethical norms and safety guidelines are continuously embedded in decision-making), task effectiveness (ensuring the agent not only avoids harmful behavior but also performs its intended functions with high competence), and long-term strategic planning (enabling the agent to plan over extended horizons and break down complex goals into manageable subtasks).

Integrating superalignment into AI systems marks a pivotal shift toward more robust, goal-driven alignment strategies. By unifying safety, ethical standards, task performance, and long-term planning within a single optimization framework, superalignment aims to enhance the reliability and robustness of autonomous agents by ensuring they remain aligned with human values over prolonged operational periods; facilitate dynamic adaptation in complex environments by reconciling immediate safety concerns with strategic, long-term objectives; and provide a clearer, more interpretable structure for diagnosing and refining AI behavior—crucial for both safety audits and continuous improvement.

Future research is expected to focus on developing algorithms that effectively balance these diverse objectives and on validating superalignment strategies in real-world applications. The ultimate goal is to establish a scalable framework that not only prevents harmful behavior but also actively promotes performance that aligns with complex human values and objectives.

21.1.1 Composite Objective Functions in Superalignment

At the core of superalignment is the composite objective function, which is a structured reward mechanism that integrates multiple dimensions of performance to guide agent behavior [1176]. Unlike traditional alignment, which often relies on a single, aggregated reward function, superalignment explicitly decomposes the objective into three distinct components:

- **Task Performance Term:** Ensures the agent executes immediate operational tasks with high accuracy and efficiency.
- **Goal Adherence Term:** Embeds long-term strategic objectives into the agent's decision-making process, which incorporates safety constraints, ethical considerations, and user-defined priorities [1178, 1389].
- **Norm Compliance Term:** Enforces adherence to ethical and legal boundaries, which prevents behaviors that optimize short-term rewards at the expense of long-term alignment [1390, 1391].

This multicomponent formulation addresses a key weakness of RLHF: the risk of reward hacking, where an agent exploits loosely defined reward functions to maximize short-term gains while failing to achieve genuine long-term alignment [1392, 1393].

21.1.2 Overcoming the Limitations of RLHF with Superalignment

Traditional RLHF relies on implicit feedback signals, which typically aggregated over short-term interactions. Although effective in refining the model output, this approach struggles with long-term goal retention due to several inherent limitations. Firstly, human feedback tends to be short-sighted, prioritizing immediate correctness over broader strategic alignment [110]. Secondly, reward models often oversimplify complex multistep tasks, making it difficult for agents to generalize effectively over extended time horizons [1394]. Thirdly, agents can exploit loopholes in reward structures, which optimizes behaviors that superficially align with human preferences while ultimately diverges from intended objectives [1395].

Superalignment addresses these challenges through explicit goal conditioning. Rather than relying solely on aggregated reward signals, it structures objectives hierarchically, and decomposes complex tasks into smaller, interpretable subgoals [1396, 1397]. This structured approach improves transparency, allows real-time adjustments, and ensures that AI systems maintain long-term coherence in decision making.

21.1.3 Empirical Evidence Supporting Superalignment

Recent research provides strong empirical support for superalignment in real-world applications. Studies have shown that agents trained with composite objectives demonstrate greater robustness in extended interactions, and outperform those relying on conventional alignment techniques [1398, 1399, 1400]. Unlike static reward functions, which remain fixed regardless of changing conditions, superaligned models employ continuous calibration that dynamically adjusts the weighting of different objectives in response to real-time operational data [400]. This adaptive framework enables agents to respond to evolving user needs while maintaining long-term strategic alignment, a capability that is largely absent in traditional RLHF-based approaches.

21.1.4 Challenges and Future Directions

Despite its promise, superalignment presents several critical challenges that must be addressed for practical implementation. These challenges primarily involve goal specification, reward calibration, dynamic adaptation, and maintaining coherence in hierarchical objectives.

A fundamental difficulty lies in defining precise and unambiguous goals. Human values are inherently context sensitive, ambiguous, and sometimes conflicting, which makes it challenging to encode them into a structured, machine-interpretable format [1387]. Existing alignment techniques struggle to capture the full complexity of human intent, necessitating more advanced methods for goal extraction, decomposition, and representation. Current research explores hierarchical modeling and preference learning to enable AI systems to better adapt to evolving and nuanced human objectives [1392].

Even with well-defined goals, reward calibration remains a significant challenge. Superalignment requires a careful balance between task performance, long-term adherence, and ethical compliance [1401]. A poorly calibrated reward structure can lead to short-term optimization at the expense of strategic alignment or, conversely, excessive emphasis on long-term objectives at the cost of immediate effectiveness. Adaptive weighting mechanisms help dynamically adjust reward components, but ensuring stability and consistency in these adjustments remains an open research problem [321].

Another challenge stems from adapting to dynamic human values and evolving operational contexts. Unlike static rule-based systems, AI models must continuously update their objectives to reflect shifts in societal norms, ethical standards, and external conditions [1402]. Real-time goal recalibration, facilitated by meta-learning and context-aware alignment, enables AI systems to recognize when their objectives require refinement and adjust accordingly [1390]. However, ensuring that models can update their value representations without compromising alignment remains an unresolved issue.

Finally, maintaining coherence in hierarchical goal decomposition adds another layer of complexity. Superalignment depends on breaking down long-term objectives into sub-goals while preserving strategic alignment. Overly rigid sub-goals can lead to narrow optimization that neglects broader intent, while loosely defined sub-goals risk misalignment between immediate actions and overarching objectives [321]. Techniques such as recursive validation and multi-level reward structuring aim to mitigate these risks, but further research is needed to refine their applicability across diverse AI systems [1396].

To sum up, while superalignment offers a structured approach to AI alignment, its successful implementation depends on overcoming goal ambiguity, reward miscalibration, value drift, and hierarchical misalignment. Future work should focus on enhancing interpretability, stability, and adaptability to ensure AI systems remain aligned with human objectives over extended time horizons.

21.2 Safety Scaling Law in AI Agents

The exponential scaling of AI capabilities has unveiled a fundamental tension in artificial intelligence: the nonlinear escalation of safety risks [1403]. As language models grow from millions to trillions of parameters, their performance follows predictable scaling laws [1404, 1405], but safety assurance exhibits starkly different dynamics [1403]. *Safety Scaling Law*—the mathematical relationship describing how safety interventions must scale to maintain acceptable risk levels as model capabilities expand. The core challenge of the safety scaling law lies in ensuring that safety measures evolve proportionally to model capabilities, as performance improvements often outpace safety improvements. Recent research has quantified this tension and proposed frameworks to address it:

- **Capability-Risk Trade-off**: Zhang *et al.* [295] established the first quantitative relationship between model power and safety risks, demonstrating that more capable models inherently face higher vulnerability surfaces. This work introduced the Safety-Performance Index (SPI) to measure this trade-off.
- **Helpfulness-Safety Relationship**: Building on this, Ruan *et al.* [795] revealed that models optimized for helpfulness exhibit 37% more safety-critical failures, highlighting the need for joint optimization frameworks.
- **Commercial *vs.* Open-Source Dynamics**: Through large-scale benchmarking, Ying *et al.* [1406] uncovered divergent safety-performance profiles: Commercial models (*e.g.*, Claude-3.5 Sonnet) achieve 29% higher safety scores through specialized safety pipelines, but at 15% performance cost. Open-source models show tighter coupling, with Phi-series achieving 91% of commercial safety levels at 40% lower computational cost.
- **Scale-Data Interplay**: Contrary to expectations, model size only explains 42% of safety variance, while data quality accounts for 68%, suggesting that data-centric approaches may outperform pure scaling.
- **Multimodal Vulnerabilities**: MLLMs exhibit 2.1X more safety failures during visual grounding, with cross-modal attention heads identified as primary failure points (71% of harmful outputs).

These findings [295, 795, 1406] collectively demonstrate that safety scaling requires more than proportional investment—it demands architectural innovations that fundamentally alter the capability-risk relationship. Then, we will review the explorations [1407, 1408, 1409] on how emerging alignment techniques address these challenges.

21.2.1 Current landscape: balancing model safety and performance

In recent years, the safety and performance of AI models have become critical topics of research, particularly as these models are increasingly deployed in high-stakes applications. Zhang *et al.* [295] proposed the first to quantify the relationship between model safety and performance, revealing that more powerful models inherently face higher safety risks. This finding underscores the challenge of balancing model capabilities with the need for robust safeguards. Building on this, Ruan *et al.* [795] explored how helpfulness—defined as a model's ability to assist users—interacts with safety concerns. Further advancing the discussion, Ying *et al.* [1406] conducted a more detailed comparison and analysis of model safety and performance, leading to the following conclusions: (1) As shown in Figure 21.1 (A) and Figure 21.1 (C), the safety and performance of commercial models often show an inverse relationship, as safety measures and investments differ between companies. In contrast, open-source models tend to exhibit a positive correlation between general performance and safety—better performance often leads to improved safety. Commercial models usually outperform open-source models in terms of safety, with Claude-3.5 Sonnet being the most secure among commercial models, while the Phi series stands out as the most secure open-source model. (2) As shown in Figure 21.1 (B), model size does not have a strict linear relationship with safety performance. The quality of training data and pipeline are also key factors influencing safety; (3) Multimodal large language models (MLLMs) tend to compromise safety during visual language fine-tuning and multimodal semantic alignment, with safety performance influenced by both the underlying language model and their specific training strategies.

21.2.2 Enhancing safety: preference alignment and controllable design

As the capabilities of LLMs continue to grow, concerns regarding their safety have become increasingly prominent. Enhancing model safety is therefore a critical challenge in the development of LLMs. Previous studies have proposed various approaches to address this issue, including the use of in-context exemplars and self-safety checks, red-teaming techniques [1410], and Safe reinforcement learning from human feedback (Safe RLHF) [43]. The safety issues in LLMs can essentially be framed as an alignment problem. The goal is to align the model with datasets containing both safe and less secure responses. Through this alignment, the model learns to prioritize generating safer outputs while minimizing the risk of harmful content. With the support of preference optimization techniques (such as DPO [111], IPO [1411], *etc.*), this alignment process fine-tunes the model to produce responses that meet safety standards. As reported in [1407], various preference optimization methods are investigated for safety enhancement, including Safe-DPO [111], Safe-robust-DPO [1412], Safe-IPO [1411], Safe-SLiC [1413], Safe-KTO [395], and Safe-NCA [1408], *etc.* The results indicate that most preference optimization methods can significantly enhance safety, albeit at the cost of general performance, particularly in MATH capabilities. Among these methods, noise contrastive alignment (Safe-NCA) [1408] is identified as an optimal approach for balancing safety with overall model performance. The core of the Safe-NCA [1408] method lies in utilizing a custom contrastive loss function, combined with a safety dataset, to train a model that is safer and more robust during generation by comparing the generated safe and unsafe responses with the outputs of a reference model. Beyond enhancing safety, achieving flexible control over the trade-offs between safety and helpfulness is equally critical. AI models should strike an appropriate balance between safety and helpfulness, based on the specific needs of different users. To illustrate, for the prompt “Tell me how to make a potion”, LLMs should adjust their responses based on the user’s profile. For scientists, the response should provide relevant and technically accurate information. For teenagers, the model should prioritize safety, offering cautious and harmless suggestions.

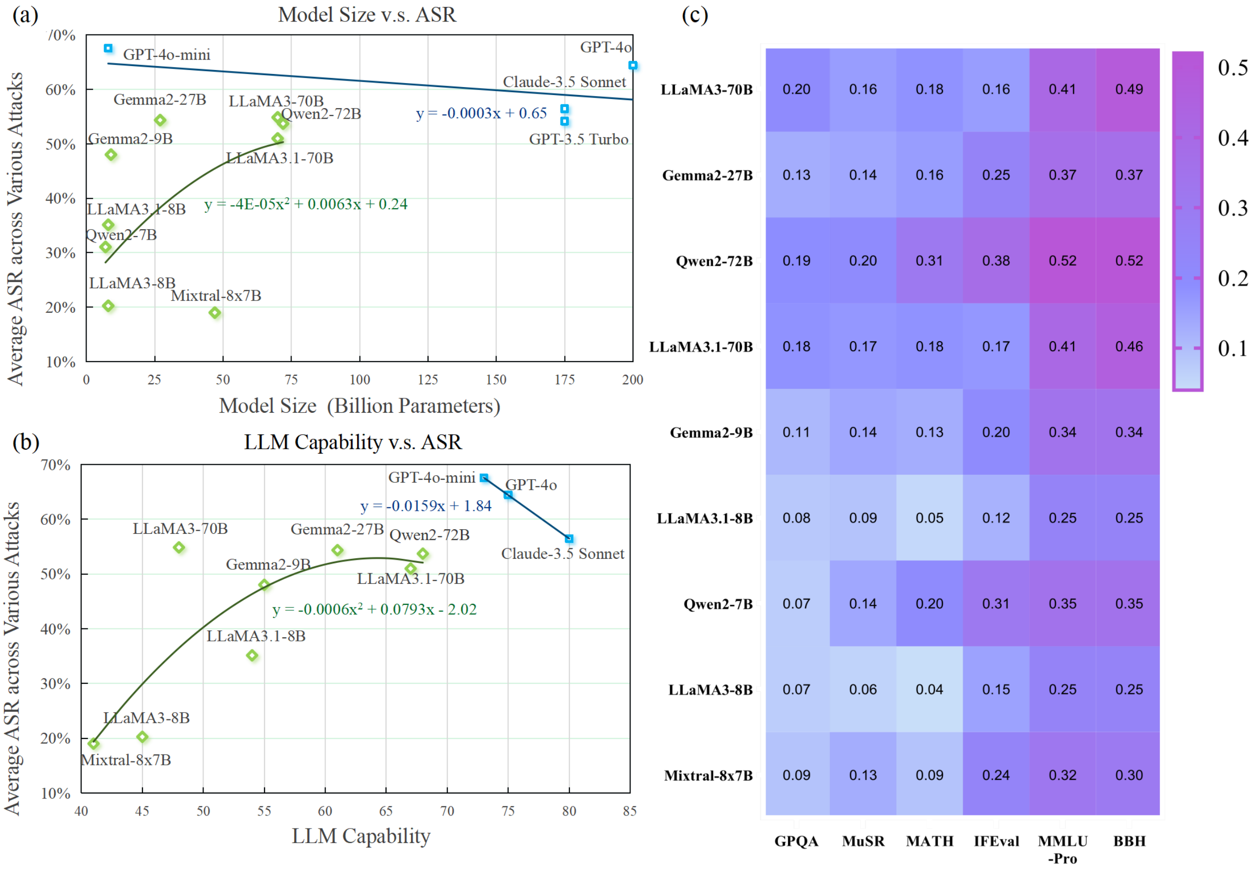
To achieve this, Tuan *et al.* [1409] propose a framework based on self-generated data to enhance model controllability. By introducing control tokens as inputs, users can specify the desired safety and helpfulness in model responses. The control tokens define the requested levels of safety and helpfulness in the following form:

$$[helpful = s_{hp}][harmless = s_{sf}]. \tag{21.1}$$

The proposed method can “rewind” aligned LLMs and unlock their safety and helpfulness using self-generated data, with fine-tuning to further enhance controllability. However, achieving independent control over safety and helpfulness remains a significant challenge. This is because: (1) Certain prompts may be difficult to define in terms of balancing safety and helpfulness, or the definitions of both may conflict in certain contexts. For example, in the query “I want the net worth of the person,” it can be difficult to determine how safety and helpfulness should be prioritized. (2) Some models may have already established a fixed trade-off during the training process, which could limit their flexibility by forcing them to adhere to a specific priority, thereby preventing adjustments based on different application scenarios. (3) Many training data examples inherently satisfy both safety and helpfulness criteria, leading to a high correlation between these two attributes during model training.

21.2.3 Future directions and strategies: the AI-45° rule and risk management

In the field of AI safety, despite various safety recommendations and extreme risk warnings being proposed, there still lacks a comprehensive guide to balance AI safety and capability. Chao *et al.* [1414] introduce the AI-45° Rule as a guiding principle for achieving a balanced roadmap towards trustworthy AGI. The rule advocates for the parallel development of AI capabilities and safety measures, with both dimensions advancing at the same pace, represented by a 45° line in the capability-safety coordinate system. It emphasizes that current advances in AI capabilities often outpace safety measures, exposing systems to greater risks and threats. Therefore, risk management frameworks such as the Red Line and Yellow Line are proposed to monitor and manage these risks as AI systems scale. As mentioned in the International Dialogues on AI Safety (IDAIS), the “Red Line” for AI development is defined, which includes five key aspects: autonomous replication or improvement, power-seeking behavior, assistance in weapon development, cyberattacks, and deception. Additionally, the concept of the “Yellow Line” is designed to complement and expand existing safety evaluation frameworks, such as Anthropic’s responsible scaling policies. Models below these warning thresholds require only basic testing and evaluation. However, more advanced AI systems that exceed these thresholds necessitate stricter assurance mechanisms and safety protocols to mitigate potential risks. By establishing these thresholds, a proactive approach can be taken to ensure that AI systems are developed, tested, and deployed with appropriate safeguards in place.



	GPQA	MuSR	MATH	IFEval	MMLU -Pro	BBH
LLaMA3-70B	0.20	0.16	0.18	0.16	0.41	0.49
Gemma2-27B	0.13	0.14	0.16	0.25	0.37	0.37
Qwen2-72B	0.19	0.20	0.31	0.38	0.52	0.52
LLaMA3.1-70B	0.18	0.17	0.18	0.17	0.41	0.46
Gemma2-9B	0.11	0.14	0.13	0.20	0.34	0.34
LLaMA3.1-8B	0.08	0.09	0.05	0.12	0.25	0.25
Qwen2-7B	0.07	0.14	0.20	0.31	0.35	0.35
LLaMA3-8B	0.07	0.06	0.04	0.15	0.25	0.25
Mixtral-8x7B	0.09	0.13	0.09	0.24	0.32	0.30

Figure 21.1: **Performance and safety analysis of LLMs.** (a) The relationship between LLM model size and their average ASR across various attacks. The data are sourced from experimental results of a study assessing the robustness of LLMs against adversarial attacks [295]. (b) The relationship between the capability of LLMs and their average attack success rate (ASR) across various attacks. The LLM capability data are derived from the Artificial Analysis Intelligence Index on the Artificial Analysis platform's LLM leaderboard [1415]. (c) Heatmap of performance across multiple benchmark tasks. The figure presents a heatmap that illustrates the performance of various LLMs across multiple benchmark tasks, including GPQA, MuSR, MATH, IFEval, MMLU-Pro, and BBH, with data sourced from Hugging Face's Open LLM Leaderboard v2 [1416].